

Deepfake Detection Using Hybrid Deep Learning with Spatial-Frequency-Temporal Feature Analysis

Indore Tanuja Sunil¹, Bhosale Sachin Bajirao², Bangar Abhishek Popat³

^{1,2,3}Department of Computer Engineering / Jaihind College of Engineering / SPPU, Pune, India

Abstract- The rapid advancement of artificial intelligence and deep learning has made it increasingly easy to create realistic manipulated images and videos, commonly known as deepfakes. Such manipulated content can create serious risks to digital security, privacy, identity protection, and the reliability of online information. Existing deepfake detection methods mainly focus on spatial features, frequency artifacts, or temporal inconsistencies, but their performance may decrease when dealing with compressed videos, unseen manipulation techniques, and cross-dataset conditions. This paper proposes a Hybrid Spatial-Frequency-Temporal Siamese Deep Learning Framework for robust deepfake detection. The proposed approach combines spatial features extracted from RGB facial frames using a convolutional neural network, frequency-based features obtained using Error Level Analysis (ELA) and frequency analysis, and temporal features learned from sequential video frames using a bidirectional recurrent model. A Siamese learning mechanism is incorporated to improve the separation between authentic and manipulated representations, inspired by recent forgery verification research. The extracted features are fused and classified into Real and DeepFake categories. Grad-CAM is also considered to provide visual explanations of the model's decisions. The proposed framework aims to improve detection robustness, generalization, and interpretability under different manipulation and compression conditions.

Keywords— DeepFake Detection; Deep Learning; Siamese Network; Spatial Features; Frequency Analysis; Temporal Features; CNN; BiLSTM; Error Level Analysis; Video Forgery Detection

I. INTRODUCTION

The rapid growth of Artificial Intelligence (AI) and Deep Learning has made it easier to create realistic manipulated images and videos known as DeepFakes. Such content can be difficult to distinguish from genuine media and may cause misinformation, identity theft, privacy violations, and cybercrime [1], [2].

DeepFake detection uses techniques such as Error Level Analysis (ELA), CNNs, attention mechanisms, and recurrent networks to identify manipulated content [2], [3]. However, existing methods may face challenges with video compression, unseen manipulation techniques, and different datasets [3], [4].

Siamese deep learning has also been used for offline signature verification to distinguish genuine and forged samples [5]. This similarity-learning concept provides useful motivation for DeepFake authenticity detection.

This study proposes a Hybrid Spatial-Frequency-Temporal Siamese Deep Learning Framework that combines RGB spatial features, ELA and frequency features, and temporal features from video frames. The proposed system aims to improve Deep-Fake detection accuracy, robustness, and explainability.

II. REVIEW METHODOLOGY

A systematic review of five selected research papers was conducted to understand the recent developments in DeepFake and digital forgery detection. The selected studies were analyzed based on their detection techniques, deep learning architectures, datasets, feature extraction methods, performance, and identified limitations. The reviewed approaches include hybrid CNN models, Error Level Analysis (ELA), EfficientNet with bidirectional recurrent networks, Multi-Head Attention, and Siamese deep learning [1]–[5].

The reviewed studies show that CNN-based methods are effective for extracting spatial features,

while ELA and texture-based methods help identify manipulation traces. Recurrent and attention-based models improve the analysis of temporal and important facial features in videos [2]–[4]. The Siamese learning approach used for signature verification demonstrates the usefulness of similarity-based learning for distinguishing genuine and forged samples [5].

From the comparative review, several challenges were identified, including poor cross-dataset generalization, video compression effects, unseen manipulation techniques, computational complexity, and difficulty in identifying subtle forgery patterns. These limitations indicate the need for a more robust and integrated detection framework.

Based on these findings, the proposed study combines spatial, frequency, temporal, and Siamese features to improve the reliability and robustness of DeepFake detection

also reported an inference time of less than 100 milliseconds, indicating its suitability for practical applications. Although the study focuses on handwritten signatures rather than DeepFake videos, its use of spatial-sequential feature learning and similarity-based verification provides useful methodological motivation for the proposed DeepFake detection framework.

2. ELA-Based DeepFake Detection

Rafique et al. [2] proposed a DeepFake detection and classification approach using Error Level Analysis (ELA) and deep learning. ELA is used to identify differences in image compression levels that may indicate regions affected by digital manipulation. The approach combines ELA-based preprocessing with deep learning models for extracting meaningful features from manipulated facial images. The study investigates CNN-based feature extraction together with machine learning classifiers such as SVM and KNN.

The research demonstrates that manipulation traces that are difficult to observe visually can be identified through error-level information. Different pretrained CNN architectures, including GoogLeNet, ResNet18, and SqueezeNet, were investigated for feature extraction and classification. The study reported strong detection performance, with the best reported accuracy reaching approximately 89.5% using a Res-Net-based approach with KNN. However, the research also highlights challenges associated with image quality, compression, noise, and generalization to different types of manipulated content. Therefore, ELA provides an important frequency/error-domain feature that can complement conventional RGB-based spatial features in a more comprehensive DeepFake detection system.

III. LITERATURE REVIEW

1. Signature Verification Using Siamese Learning

Basnet et al. [5] presented a Bidirectional Recurrent Convolutional Siamese Network (BRCSN) for offline, writer-independent signature verification. The main objective of the study was to distinguish genuine signatures from forged signatures without requiring writer-specific training. The proposed framework combines Convolutional Neural Networks (CNNs) for extracting spatial features with Bidirectional Gated Recurrent Units (Bi-GRUs) for learning sequential patterns. A triplet loss function is used to learn a discriminative feature representation by reducing the distance between genuine samples while increasing the distance between genuine and forged samples.

The study demonstrates that Siamese-based learning is useful for forgery detection because it focuses on the similarity and differences between samples rather than relying only on conventional classification. The model was evaluated on benchmark multilingual signature datasets and showed good classification performance with low false acceptance and rejection rates. The authors

3. Deep Learning-Based Video Detection

Several recent studies have applied advanced deep learning techniques for detecting DeepFake videos. The study reported in [1] investigates DeepFake detection using deep learning and discusses hybrid CNN architectures such as Inception, ResNet v2, and Xception for extracting frame-level features. The method was evaluated using the DeepFake

Detection Challenge (DFDC) dataset, demonstrating the effectiveness of deep learning for identifying manipulated visual content. However, frame-level approaches may not completely capture temporal inconsistencies occurring across consecutive video frames.

Agrawal and Haneef [3] proposed PAttSBiL, an integrated deep learning framework for DeepFake video detection. The system converts videos into frames and uses YOLOv5 for face localization. Facial landmarks and eye-related features such as the Eye Aspect Ratio (EAR) are also considered. EfficientNet-based feature extraction is combined with a Pelican Attention Stacked Bidirectional LSTM to capture important spatial and temporal information. The framework was evaluated using datasets such as FaceForensics++ and Celeb-DF(v2), achieving high detection performance. The study also emphasizes that generalization across different DeepFake generation techniques remains a significant challenge.

Sekar et al. [4] proposed an optimal deep learning model with Multi-Head Attention-based feature extraction for DeepFake detection. Their framework consists of face detection, preprocessing, texture feature extraction, spatial feature extraction, and classification. A Butterfly Optimized Gabor Filter is used to obtain texture information, while ResNet-50 with Multi-Head Attention extracts important spatial features. An optimized LSTM is then used for classification. The method was evaluated on multiple benchmark datasets, including FaceForensics++, DFDC, Celebrity Deep-Fake, and WildDeepFake. The study shows that attention-based feature learning can improve DeepFake detection; however, cross-dataset performance continues to be a challenging research problem.

Overall, the reviewed studies demonstrate that CNNs are effective for spatial feature extraction, ELA can provide useful manipulation-related information, and recurrent or attention-based models can capture temporal dependencies in videos [1]–[4]. The signature verification study further demonstrates the potential of Siamese learning for distinguishing genuine and forged samples [5]. However, the

existing approaches generally focus on individual feature domains or specific architectures. Therefore, there is a need for an integrated framework that combines spatial, frequency, temporal, and similarity-based features to improve the robustness of DeepFake detection, particularly against compressed videos, unseen manipulation techniques, and cross-dataset variations.

4. Multi-Head Attention DeepFake Detection

Sekar et al. [4] proposed an optimal deep learning approach for DeepFake detection using an attention-based feature extraction scheme. The proposed framework consists of face detection, preprocessing, texture feature extraction, spatial feature extraction, and classification. A Butterfly Optimized Gabor Filter is used to extract important texture features from the detected face regions. These features help in identifying local patterns and possible manipulation traces present in DeepFake images and videos.

For spatial feature extraction, the study uses ResNet-50 with Multi-Head Attention. The Multi-Head Attention mechanism helps the model focus on important regions and learn global relationships within the facial features. After spatial feature extraction, an optimized LSTM model is used for classification of the input as real or fake. The proposed method was evaluated using multiple datasets, including FaceForensics++, DFDC, Celebrity DeepFake, and WildDeepFake.

The study demonstrates that combining texture features, deep residual networks, attention mechanisms, and recurrent learning can provide effective DeepFake detection. However, the authors identify cross-dataset generalization as an important challenge, since a model trained on one dataset may show reduced performance when tested on different datasets or unseen manipulation techniques [4].

5. Hybrid Deep Learning-Based DeepFake Detection

The fifth study [1] investigates DeepFake detection using deep learning techniques and focuses on identifying manipulated visual content

automatically. The study discusses the increasing difficulty of distinguishing real and fake media and reviews different deep learning approaches used for DeepFake detection. It highlights the importance of extracting meaningful visual features from video frames for reliable classification.

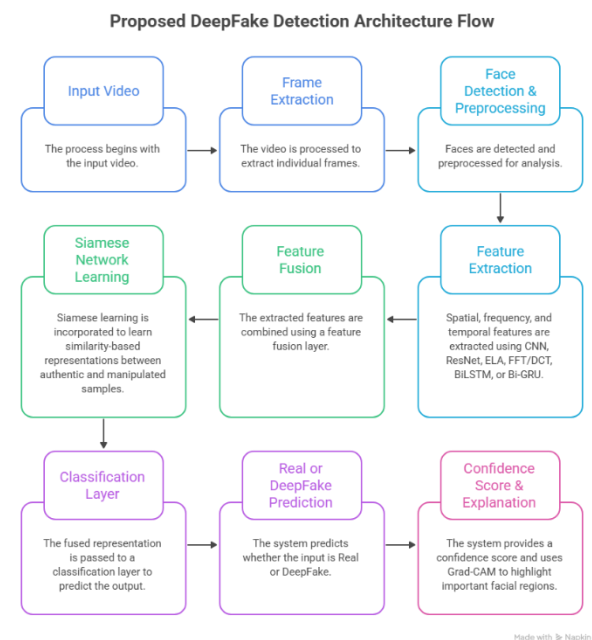
A hybrid CNN architecture combining Inception, ResNet v2, and Xception is used for frame-level feature extraction. These networks provide complementary feature representations that help identify subtle visual patterns associated with manipulated content. The approach was evaluated using the DeepFake Detection Challenge (DFDC) dataset, and the reported results showed high precision, recall, and F1-score, indicating the effectiveness of the hybrid deep learning approach [1].

However, frame-level analysis mainly focuses on individual images and may not fully capture temporal inconsistencies between consecutive frames. This indicates the need for combining spatial features with temporal information and other manipulation-sensitive features. Therefore, the study provides a strong foundation for developing a more comprehensive DeepFake detection framework

IV. COMPARATIVE ANALYSIS

The reviewed studies use different deep learning techniques for detecting DeepFake and other forms of digital forgery. Their comparison is presented below

Author, Year	Objective	Method / Technology	Major Reported Findings	Limitations
Ikram et al., 2023 [1]	DeepFake image/video detection	Hybrid CNN using Inception, ResNet v2 and Xception	High precision, recall and F1-score on DFDC dataset	Mainly frame-level analysis; temporal information is limited
Rafique et al., 2023 [2]	DeepFake detection and classification	Error Level Analysis (ELA) + CNN + SVM/KNN	Best reported accuracy of about 89.5% using ResNet + KNN	Sensitive to compression, blur and unseen content
Agrawal & Haneef, 2026 [3]	DeepFake video detection	YOLOv5 + facial landmarks + EfficientNet + PAttSBiL	97.41% accuracy on FaceForensics++ and 97.87% on Celeb-DF(v2)	Cross-dataset generalization remains challenging
Sekar et al., 2025 [4]	Accurate DeepFake detection	Gabor Filter + ResNet-50 + Multi-Head Attention + LSTM	Up to 99.78% accuracy; cross-validation reached 94.65%	Cross-dataset performance is still challenging
Basnet et al., 2025 [5]	Offline signature forgery verification	CNN + Bi-GRU + Siamese Network + Triplet Loss	Good classification performance with low false acceptance/rejection; inference under 100 ms	Focused on signature forgery, not DeepFake detection



V. PROPOSED CONCEPTUAL ARCHITECTURE

Layer 1 receives the input video and extracts individual frames from the video. Layer 2 performs face detection and preprocessing, including face cropping, resizing and normalization. Layer 3 performs multi-domain feature extraction using three parallel branches. The spatial branch uses CNN or ResNet to extract facial and visual features from RGB frames, while the frequency branch applies ELA and FFT/DCT followed by CNN to identify manipulation-related traces. The temporal branch uses BiLSTM or Bi-GRU to analyze sequential frames and capture frame-to-frame inconsistencies and unnatural facial movements. Layer 4 fuses the spatial, frequency and temporal features into a combined feature representation. Layer 5 applies Siamese similarity learning to compare the extracted representation with authentic and manipulated samples and learn discriminative representations using contrastive or triplet-based learning. Layer 6 passes the learned representation to a classification layer to predict whether the input video is REAL or DEEPFAKE and provides a confidence score. Layer 7 provides explainability using Grad-CAM by highlighting the important facial regions that influenced the model's decision. The proposed architecture therefore combines spatial, frequency,

temporal and similarity-based information to improve the robustness and interpretability of DeepFake detection.

VI. RESEARCH GAPS AND FUTURE DIRECTIONS

Gap supported by evidence	Supporting studies	Future direction
Limited cross-dataset generalization	[3],[4]	Evaluate the proposed model on multiple unseen datasets and manipulation techniques
Sensitivity to compression, blur and degraded video quality	[2]	Test compression-resilient detection using ELA and frequency-domain features
Limited temporal analysis in frame-level approaches	[1]	Integrate BiLSTM/Bi-GRU to capture temporal inconsistencies across video frames
Limited integration of spatial and frequency features	[1], [2], [4]	Develop hybrid spatial-frequency feature fusion using CNN/ResNet with ELA and FFT/DCT
Siamese learning mainly explored for forgery verification	[5]	Adapt Siamese similarity learning for DeepFake authenticity representation
High performance often reported within individual datasets	[3], [4]	Perform cross-dataset and cross-manipulation validation to measure robustness

Limited explainability of DeepFake predictions	[3], [4]	Integrate Grad-CAM to highlight facial regions influencing the prediction
Confidence and robustness under real-world conditions require further investigation	[1]–[5]	Provide confidence scores and evaluate performance under different video-quality conditions

Directions still requiring investigation include lightweight real-time deployment, evaluation against newly generated DeepFake techniques, robustness to severe compression and noise, and the development of explainable detection systems that can provide both prediction confidence and visual evidence for the final decision

VII. CONCLUSION

The reviewed studies demonstrate that DeepFake detection has advanced significantly through the use of CNNs, Error Level Analysis, recurrent networks, attention mechanisms, and Siamese learning [1]–[5]. However, challenges such as cross-dataset generalization, video compression, unseen manipulation techniques, limited temporal analysis, and lack of explainability still remain. The comparative analysis shows that individual techniques can achieve strong performance, but they may not be sufficient under diverse real-world conditions.

To address these gaps, this study proposes a Hybrid Spatial-Frequency-Temporal Siamese Deep Learning Framework for Robust DeepFake Detection. The proposed architecture combines CNN/ResNet-based spatial features, ELA with FFT/DCT-based frequency features, and BiLSTM/Bi-GRU-based temporal features. These features are fused and further processed using Siamese similarity learning before classification into Real or DeepFake. Confidence scoring and Grad-CAM are also

proposed to improve the interpretability of the detection results.

The proposed framework provides a direction toward a more robust, explainable, and generalized DeepFake detection system. Future implementation and evaluation on multiple datasets, unseen manipulation methods, and different compression conditions will be required to validate its effectiveness and practical applicability.

REFERENCES

1. M. M. Rafique, Z. H. Qaiser, M. Fuzail, N. Aslam, and M. S. Maqbool, "Implementation of Efficient Deep Fake Detection Technique on Videos Dataset Using Deep Learning Method," *Journal of Computing & Biomedical Informatics*, vol. 5, no. 1, 2023. DOI: 10.56979/501/2023
2. R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," *Scientific Reports*, vol. 13, article 7422, 2023. DOI: 10.1038/s41598-023-34629-3
3. D. R. Agrawal and F. Haneef, "PAttSBiL: An Efficient Deep Fake Video Detection Using Integrated Deep Learning Methodologies," *Multimedia Tools and Applications*, vol. 85, 2026. DOI: 10.1007/s11042-026-21854-x
4. R. Raja Sekar, T. Dhiliphan Rajkumar, and K. R. Anne, "Deep Fake Detection Using an Optimal Deep Learning Model with Multi Head Attention-Based Feature Extraction Scheme," *The Visual Computer*, vol. 41, pp. 2783–2800, 2025. DOI: 10.1007/s00371-024-03567-0
5. S. Basnet, D. R. Dongol, and S. Shrestha, "Advancing Offline Signature Verification with Bidirectional Siamese Deep Learning: A Writer-Independent Approach," *Journal of Engineering Issues and Solutions*, vol. 4, no. 1, 2025. DOI: 10.3126/joeis.v4i1.81613
6. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020. DOI: 10.1016/j.inffus.2020.06.014

7. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11, 2019. DOI: 10.1109/ICCV.2019.00009
8. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deep-Fake Detection Challenge (DFDC) Dataset," arXiv preprint, 2020. DOI: 10.48550/arXiv.2006.07397
9. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3207–3216, 2020. DOI: 10.1109/CVPR42600.2020.00327
10. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019. DOI: 10.1109/ICASSP.2019.8682602
11. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018. DOI: 10.1109/WIFS.2018.8630761
12. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018. DOI: 10.1109/AVSS.2018.8639163